Laboratory Animal Science

Vol 41, No 1
January 1991

SPECIAL TOPIC OVERVIEW

Appropriate Animal Numbers in Biomedical Research in Light of Animal Welfare Considerations

MD Mann, DA Crouse and ED Prentice

Over the last decade, public awareness of the use of animals in biomedical research has generally increased, whereas public attitudes toward such research have not necessarily changed. Most people still recognize the potential benefits of animal research and believe that these benefits outweigh the costs. At the same time, however, the public is concerned that animals are used both humanely and wisely, e.g., wisely in terms of conserving natural resources. Generally, scientists reflect the attitudes of the population and there has been a concerted effort to reduce the number of animals used in research.

Scientific granting agencies, oversight bodies such as the U.S. Department of Agriculture (USDA), and the scientific community have all endorsed the use of Russell and Burch's three Rs (1). Russell and Burch recommended that efforts be made to Replace animals with nonanimal models, Reduce the number of animals used in research, and Refine the techniques of research to reduce animal suffering. Clearly, the intent of the Animal Welfare Act of 1985 (2) and the rules resulting from that law are designed to encourage the research community to implement these three principles. This is reflected in the following passage from the Public Health Service (PHS) *Policy on Humane Care and Use of Laboratory Animals* (3): "The animals selected should be of an appropriate species and quality and the minimum number required to obtain valid results. Methods such as mathematical models, computer simulation, and *in vitro* biological systems should be considered." The USDA animal welfare regulations (4) also reflect this concern: "A proposal to conduct an activity involving animals, or to make a significant change in an ongoing activity involving animals, must contain...A rationale for involving animals, and for the appropriateness of the species and numbers of animals to be used..."

Consistent with the national trend, the number of animals used in research at the University of Nebraska Medical Center has declined dramatically since 1979. This general downward trend in animal use is not simply the result of animal welfare legislation because its beginning clearly precedes the Animal Welfare Act of 1985. Declining usage also could be the result of increased cost of animals and their care, but increasing grant funds should have partly offset that effect. A number of factors are probably responsible for this trend, among them increased concern among scientists for animal welfare, increased cost of purchase and maintenance of animals, decreased funding of animal-related research from Federal sources, increased oversight by institutional committees, and increased emphasis on molecular approaches and biotechnology which often require fewer animals. It is not possible to rank-order these factors by importance. Clearly, the reasons for the decline in the number of animals used are multiple.

Both the PHS and USDA have charged the Institutional Animal Care and Use Committee (IACUC)[1] with ensuring appropriate and humane animal care and use. It is difficult to see how any committee could offer such assurance without considering the appropriateness of animal numbers. Five years of experience in making decisions regarding the appropriateness of animal numbers has convinced us that this is a complicated issue, an issue worthy of examination in light of animal welfare considerations. A number of factors enter into any considerations of the appropriate number of animals to be used in a research project. Statistical, economic, contractual and welfare issues all play some role in such considerations. Investigators and committees must use all of these factors in making decisions about the appropriateness of animal numbers. In this review, we will first consider the statistical aspects of experimental design and will then take up nonstatistical issues which committees encounter in the review of proposed experiments using animals.

Statistically Appropriate Numbers of Animals

The decision about the number of animals to use in an experiment often can be made on statistical grounds, that is, the appropriate minimum number of animals to be

From the Department of Physiology and Biophysics (Mann) and the Department of Anatomy (Crouse:, Prentice):, University of Nebraska Medical Center:, Omaha:, NE 68198

Table 1 Sources of sample size tables and nomograms

Statistical Test	Reference
Chi Square, 2 x 2	.11
Chi Square, rxc	.15,38
F Test	.9
Log-rank Statistic	.39
Multiple Correlation	.15,40
Odds Ratios	.14,41,42
One-way Analysis of Variance (Parametric)	.13,15,17,43
Paired *t* test	.12,44
Proportions Pearson Correlation Coefficient	.9,15
Comparing Two Proportions	.10,15,45,46
Confidence Limits	.9
Sign Test	.15
t Test	.9,12,15,44,47
Two-sample McNemar Test	.48
Two-sample McNemar Test, Matched Pairs	.49
Two-way Analysis of Variance (Parametric)	.15,43

used is specifiable given the nature of the experiment and the statistical tests to be used in analyzing the data obtained. Proper use of most statistical tests requires that the decision of what test to use precede the experiment, and, in fact, the experiment must be designed to use that particular test. In reference to analysis of variance, Sokal and Rohlf (5) remind us that, "An important point about such tests is that they are designed and chosen independently of the results of the experiment. They should be planned *before* the experiment has been actually carried out and the results obtained." Many journals are now requiring that "power-based" assessment of the adequacy of the sample size be provided in papers submitted for publication (6,7,8). Such an assessment should be made before the experiment is begun.

It is beyond the scope of this paper to present the formulae for calculation of appropriate sample sizes. Calculation formulae, sample size tables and nomograms exist for a variety of statistical tests (9,10,11,12,13,14). Table 1 presents some examples of tests for which such tables exist, but this list is by no means exhaustive. Particularly, readers are referred to an excellent text by Jacob Cohen (15) which discusses power analysis in detail and presents tables for a wide range of statistical tests. Unfortunately, there are tests commonly used in experimental research for which there are not such tables, but Erb (16) suggested some strategies for applying existing tables to similar tests for which there are no tables. Use of these formulae, tables or nomograms requires knowledge of several parameters: the acceptable alpha, type I or false positive error; the acceptable beta, type II or false negative error; the smallest difference worth detecting or the effect size (15); and the variability of the control and experimental populations.

Alpha error: Alpha error is the probability of error in concluding that there is a difference between experimental and control populations, i.e., concluding that the experimental treatment of the animals has an effect when in reality there is no effect. This is the error associated with rejecting the notorious null hypothesis, the hypothesis of no difference or no effect, when it is actually true. In parametric tests, it is the area (for so-called one-tailed tests) or areas (for two-tailed tests) under the extremes of the normal curve representing the null hypothesis (H_0 True) beyond the

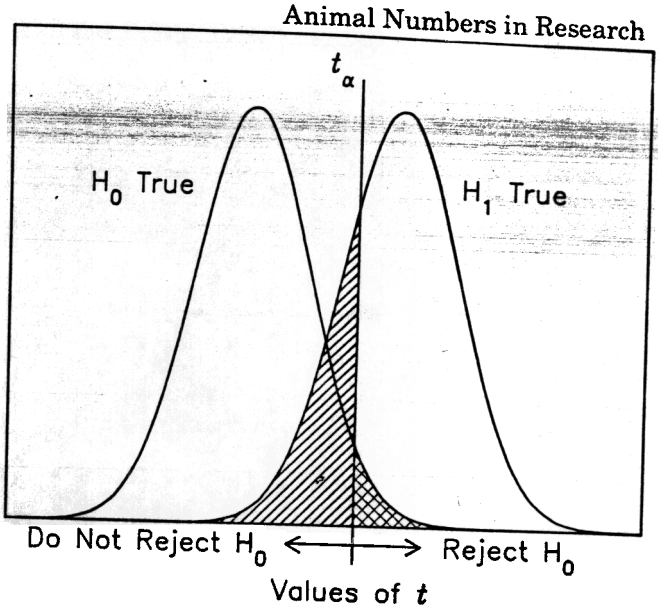

Figure 1 Sampling distribution of the *t* ratio when the null hypothesis is true (H_0 True) and when the alternative hypothesis is true (H_1 True). The criterion value of the *t* ratio is indicated by the vertical line labeled t_α. The probability associated with that criterion, α, is indicated with the area to the right of t_α under the H_0 True curve (cross hatched area), whereas the β probability is indicated by the area to the left of the t_α under the H_1 True curve (hatched area). The power of the test, 1-β, is indicated by the balance of the area under the H_1 True curve.

criterion value, in Figure 1 the cross-hatched area. Every biological population contains some variation. Random samples from that population will rarely be identical to each other. Therefore, there is a finite probability that two very different samples could be drawn from a single population. Because the samples are different, the conclusion may be drawn that they come from different populations, when, in fact, they do not.

The probability of an α error can never be zero in biological experimentation, but biologists will settle for an appropriately low, nonzero probability. This is the ubiquitous level of significance or *p* value. The acceptable α probability has been *arbitrarily* set at 0.05 or 0.01, values employed in most statistical testing. The 0.05 level indicates a willingness to incorrectly reject the null hypothesis once in 20 times when it is actually true. This probability is frequently used as follows: When the calculated probability is equal to or less than 0.05, the null hypothesis is rejected (i.e., it is concluded that the treatment had an effect), whereas, if the calculated probability is equal to or greater than 0.06, the null hypothesis is not rejected (i.e., it is concluded that the treatment had no effect). In this way, the experimenter could limit his liability of error to 1 in 20 such experiments. But consider the probability of 0.06. This corresponds to an error rate of 1 in 17 or 18 instead of 1 in 20, a small difference which, by common sense, should merit a similar interpretation. The practice of simply specifying a difference associated with a probability of 0.06 or even 0.07 as "insignificant," without specifying the actual values, is indefensible, i.e., it is preferable to specify the calculated *p* value.

Most people would agree that the smaller the pro-

Vol 41, No 1
Laboratory Animal Science
January 1991

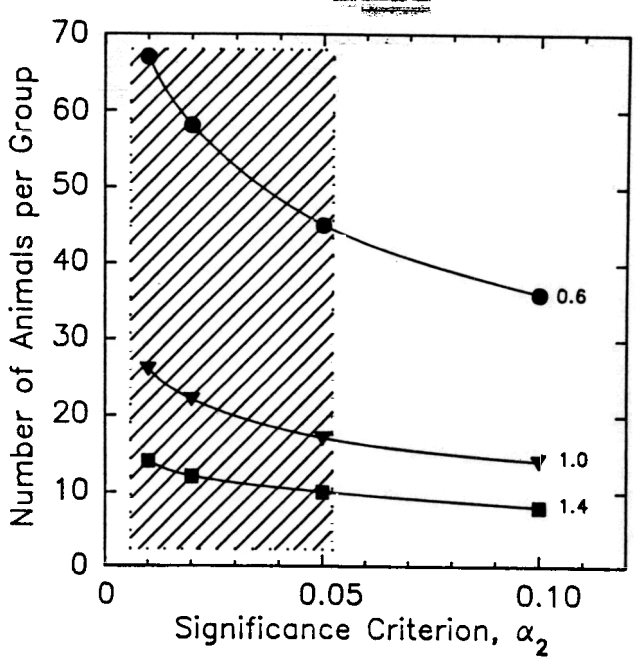

Figure 2 The relationship between the significance criterion, α_2, and the sample size. This figure was constructed for the *t* ratio for equal size samples assuming a β probability of 0.2. Because sample size depends upon the effect size, three different curves were constructed for effect sizes, 0.6, 1.0 and 1.4, where effect size, Δ, is indicated by $\Delta = (\mu_x - \mu_y)/\sigma$, with μ_x and μ_y the means of the two populations and σ the within population standard deviation, assumed to be equal in the two populations. Alpha values are for two-tailed tests. The hatched area spans the normally accepted significance criteria.

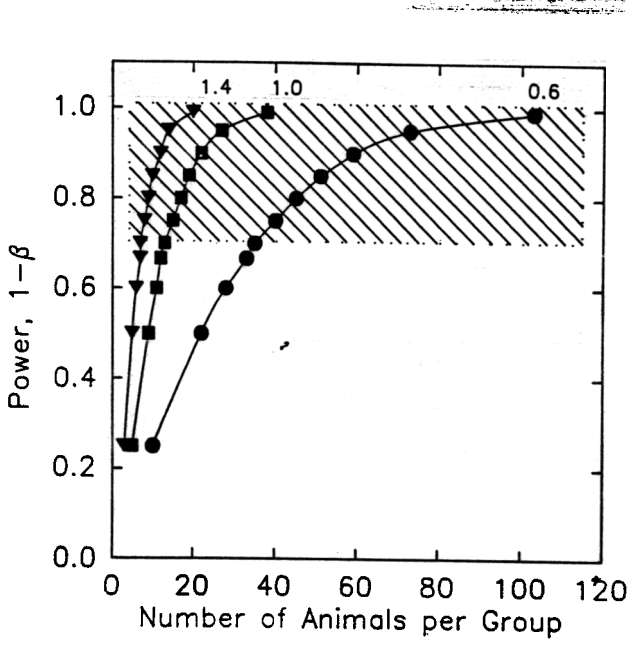

Figure 3 The relationship between the power of the test, $1-\beta$, and the sample size. The figure was constructed for the *t* ratio for equal size samples, assuming $\alpha_2 = 0.05$ (two-tailed). All other assumptions and conventions are the same as those in Figure 2. The hatched area spans the range of powers which can be considered reasonable for detection of an effect.

bability of an α error the better. There are, however, limits to the truth of this statement. In general, the smaller the α probability, the smaller the power of the experiment[2], i.e., the less likely the experiment is to find an effect of the treatment if it exists. All other things being equal, the smaller the acceptable α probability, the larger the sample size must be to detect a real effect if it exists. Even a cursory examination of any statistical table will verify this fact. In Figure 2 are plotted two-tailed significance criteria, α_2, against sample size for three different effect sizes (0.6, 1.0 and 1.4 times the standard deviation). Clearly, sample size must increase with reduced α probability. In addition, the smaller the probability (α) of accepting a false positive effect, the larger the probability of failing to reject a false null hypothesis, i.e., the larger the probability of a β error.

Beta error: The β error occurs when it is wrongly concluded that the treatment has no effect, and it is related to the power of the study, *power* = 1 - β, the probability that a difference that actually exists will be detected by the study. In parametric tests, the β error is the area under the normal curve representing the hypothesis alternative to the null hypothesis (H_1 True) beyond the criterion value; in Figure 1 it is the hatched area. The power of the test is the remainder of the area under the "H_1 True" curve. The optimal relationship between α and β depends upon the situation of the study. If committing a β error is more costly than committing an α error, as it would be if a promising treatment for cancer were incorrectly discarded, then β should be smaller than α. The loss of lives and increased suffering that might result from discarding the treatment would indeed be costly. On the other hand, if committing an α error is more costly than committing a β error, as it would be if an ineffective treatment were adopted when another, effective, treatment exists, then α should be smaller than β. Again, the loss of life and increased suffering could be significant. Animal experimenters may be tempted to make both values vanishingly small, but the additional animals required may be too high a price to pay for such certainty.

Figure 3 shows the relationship between power and sample size for the same three relative effect sizes used in Figure 2. Clearly sample size may be reduced by accepting a reduced power but, below about 0.7, even large reductions in power have relatively little effect on sample size.

Effect size: Keppel (17) has pointed out that "in most, if not all, of the experiments we conduct, the null hypothesis is *false*. That is, when we treat groups of subjects differently, they will behave differently." If these differences are small then we must be certain we have enough animals to detect them. The effect size is the actual difference between the experimental and control populations in the parameters being measured. In order to determine the appropriate sam-

ple size, the experimenter must determine the effect size he expects. In other words, he must determine "the smallest difference that is worth detecting" (16). It is possible to find even a miniscule difference statistically significant; but, though this difference is *statistically* significant, it may not be *biologically* or *clinically* significant or important. A very small improvement in the therapeutic potential of a new drug may be detectable statistically, but that small improvement may have no clinical importance if the drug is more expensive or if it has undesirable side-effects.

Cohen (15) has pointed out that "...the larger the ES [*effect size*] posited, other things (significance criterion, desired power) being equal, the smaller the sample size necessary to detect it." Conversely, the smaller the effect size, the larger the sample size must be. The appropriate sample size is that needed to detect an expected effect size, if the effect exists. Erb (16) has emphasized that a sample size larger than that needed to find the smallest worthwhile difference constitutes a waste of animals. Furthermore, a sample size smaller than that needed to detect such a difference is an inappropriate use of *all* the animals of the study.

As an example of this problem, our committee was called upon to review a protocol[3] involving use of monkeys in a study of various procedures to shorten the jaw (sagittal split-ramus osteotomy). In the initial protocol, 15 monkeys were requested to be put into 10 different experimental groups. Given the nature of the anatomical observations and measurements to be taken and the variability of such observations, no useful information could have been derived from this study. The committee, therefore, suggested that either the number of animals per group be increased or the number of groups reduced. In this case, it may be *less* wasteful of animals to use *more* of them. In fact, the committee review of this protocol led to no increase in animal usage, largely because of the projected cost, but did result in a more focused project which addressed the *most important* experimental variables.

It is difficult to say how common this error may be over the whole range of biomedical research. A hint of its commonness was given by Freiman, *et al.* (18,19) who reviewed 71 negative clinical trials. They calculated, using data presented in the reports of these trials, the power of the studies given the sample size employed and found that 67 of the studies involved insufficient patients to detect a 25% benefit of the therapy. Fifty of the studies could not have detected a 50% benefit if it existed. Freiman, *et al.* had no way of calculating the costs and benefits for each trial, but it is hard to imagine that a 50% benefit is unimportant. Even worse, most of the authors concluded that no clinically meaningful effect existed. We have no similar data for animal experiments, but we may presume that this error is not uncommon in them (20). Similar experiments involving animals would be wasteful of animals; it also may be regarded as unethical by peers (21). False negative findings have an additional effect which may be devastating to scientific progress. Such findings may prevent others from investigating the phenomena involved and certainly will affect the way others think about them.

Most psychologists (and many other investigators) are inclined not to believe any result achieved with a sample size of 5. But, both investigators and committees must be aware that, if the assumptions of a statistical test are true for such a small sample, a statistically significant result is believable. As we have seen, the difficulty comes when the result is not significant (or more correctly, when the accepted criterion value is not met).

Variability: One of the most common causes of failure to detect a real difference is excessive variability in the measurements. Control measures in experiments are important for reducing the undesirable variability. However, even if all extraneous variability were controlled, the parameters measured still would be different in different animals or subjects; individuals within a population simply differ from each other.

It is important that any sample value (mean, proportion, variance) be close to the relevant population value, i.e., it must be reliable. Reliability may be related to a number of factors, depending upon the specific statistical model being used, but it is always related to the size of the sample. For example, the most commonly used measure of reliability, the standard error of the mean, depends upon the square root of the ratio of an unbiased estimate of the population variance (s) and the sample size (n), thus, $SE = \sqrt{s^2/n}$. Similarly, the standard error of a Pearson correlation coefficient depends upon the ratio of 1 minus the square of the population coefficient (r) and the square root of the sample size minus 1, $SE = (1 - r^2)/\sqrt{n-1}$. Uniformly in statistical tests, the larger the size of the sample taken, the smaller the error and the greater the reliability of the results.

For many tests, the larger ratio of the effect size to the standard deviation, the more likely the effect is to be statistically significant. It is a common practice to increase sample size and therefore reduce the standard deviation when the effect size is small. Failure to increase the sample size results in an experiment of low power. An increase in sample size should be equivalent in both experimental and control groups. As discussed later, a large increase in only the control group can invalidate the test by violating its assumptions.

Knowing and estimating: In most cases, it is necessary to know the effect size, α and β probabilities, and variability in order to estimate minimum sample size. Alpha probabilities have been generally agreed upon by the research community, although the established values are arbitrary. Some care should be taken in too quickly accepting these values as immutable. It is possible to derive any conclusion from a study no matter what the associated probabilities. On the other hand, peer reviewers are not likely to relish conclusions of significance if the probability of an α error is too large, say >0.1. Acceptable β probabilities are not canonized. According to Neyman and Tokarska (22), a power of 0.8-0.9 should be regarded as reasonable for detection of an effect. Kraemer (23) suggested that power be 0.8. In his calculations, McCance (24) used a power of 0.7 as the lower boundary of the range of acceptable powers. Reasonable values of power are in the range of 0.7-0.9.

Frequently, the effect size and variability of the data are known from previous experiments performed either by the investigator or by others. There already may have been experiments using similar protocols, with subjects drawn from essentially the same population or from a sufficiently similar population. In these cases, rather precise values may be assigned to these variables. The experimenter may not always know these parameters, especially when making measurements of new variables. What then? One possible approach is to do a pilot study using a small sample size. If the experimenter is willing to accept the sample means and variances as adequate estimates of the population parameters, then a rough estimate of minimum sample size is possible. The minimum sample size calculated will be one that produces sufficient power to detect an effect at least as large as that obtained in the pilot study. This estimate will not be as good as that obtained when the effect size and variability are actually known, but it is better than nothing at all.

Nonstatistical Factors in Determination of Sample Size

Ethical considerations: Clearly, there are ethical implications of using too many or too few animals to detect an effect if it exists. Both are a waste of animals, though using too few may be the greater waste. There are also ethical or animal welfare considerations that influence the number of animals used, but these are considerations that have little to do with statistical analysis. For example, our committee recently was asked to review a protocol involving the use of rabbits in a study of frostbite. The investigator proposed producing frostbite in both hind limbs of a rabbit, then rewarming each limb according to a different procedure to see which would produce the least necrosis and better healing. The committee was concerned that a rabbit debilitated in this way would have great difficulty moving around in a cage and that so much damage to the leg tissues and their circulation might reduce survival of the animals. The committee recommended that the investigator consider doubling the number of animals and producing frostbite in only one leg or using the fore limbs or ears instead. These recommendations were made only for reasons of the welfare of the animal.

Frequently, such considerations lead to either a reduction in the number of treatments or an increase in the number of animals. Either outcome is justifiable on both ethical and statistical grounds. Protocols involving multiple survival surgeries (occasions for general anesthesia and surgery on the same animals separated by survival time) are reviewed carefully by the committee. These must be justified exhaustively by the investigator. Unless such justification is adequate the committee will recommend that the surgeries be done on different groups of animals, sometimes resulting in increased numbers of animals.

Economic considerations: Rowan (25) observed that "few experiments require large numbers of dogs, cats or primates, but protocols that require hundreds of rodents are common." His conclusion that "obviously" fewer large animals are used in research because they are more expensive is not accurate for many, if not most, cases. His question about the statistical reliability of large-animal experiments or the unjustifiably large numbers of rodents in experiments, while important, is meaningless in this context.

There are many reasons for choosing one species over another in experimentation. Frequently, rodents are chosen because they may be inbred, thus reducing variability by eliminating much of the genetic variation. Gall stones are easily induced in prairie dogs, but not in most other species, by a simple change in diet. The cheek pouch of the hamster is a site into which tissues can be transplanted for the study of microcirculation, the circulation through arterioles and capillaries. The squid is ideal for axonology because of its large-diameter axons. Some larger animals may be selected because something about them more closely resembles the condition in man, and so forth. Often custom is the only reason for using a particular species. For many years, cats were used in neurophysiological experiments simply because they had been used by many different investigators for many years.

That is not to say that economic considerations do not enter into the choice of species. Rats may be elected for experiments which could be done equally well in cats or dogs, because rats are less expensive to purchase and maintain. But expense should not be the only consideration in the choice of species. Experiments in screening of possible carcinogens require large numbers of animals because the expected incidence of cancer is likely to be low and it takes a long time, relative to the life-span of the species, for the cancers to appear (26). This is partly an economic issue, but it would be difficult to justify the use of large numbers of endangered primates when abundant rats would suffice.

Rowan (25) has also missed the point of the appropriate experimental units (16). For large animal experiments the proper unit of the experiment is frequently the number of measurements because a number of measurements are made on each animal. Therefore, n is the number of measurements, not the number of animals. For rodent studies, the unit of the experiment frequently is the number of animals or the number of pools of animals. Often, rodent studies involve pooling animal materials to get samples large enough to measure or large enough to produce some desired effect size. In these experiments, the number of animals required is determined by the number of experimental conditions and the number of animals per pooled sample.

Grantor imposed restrictions: When an investigator applies for a contract to perform some studies for a certain granting agency, he frequently must agree to conduct the research in a certain way. Often, the agency will specify how many animals must be used in the experiments. It sometimes appears that the numbers have not been specified for obvious or defensible reasons. It is common practice in such contracts to specify that the control groups be two to five times larger than the experimental groups. There is no statistical justification for this practice. It is possible to increase the power of the experiment by increasing the size of the control group by itself (any increase in number of animals will have this effect), but

the increase in power is only apparent!

Most statistical tests require the assumption that the variances of the experimental and control populations are homogeneous. As we have seen, increasing sample size will reduce the variance of the sample, thus increasing the ratio of the effect size to the variance. However, in comparing two samples, the *actual* power of the test is determined by the smallest sample, not the largest. Enlarging only the control group causes a violation of the assumption of homogeneity and, therefore, invalidates the statistical test.

In these cases, the committee faces a dilemma. Approval of such a protocol constitutes a waste of animals and, therefore, an abrogation of the committee's responsibilities. To refuse such a protocol means the institution will not receive the contract. It seems that it is the responsibility of the granting agency to assure that the contract contains valid specifications for the appropriate number of animals. In some cases, the appropriate number of animals is not specified by an agency, but by convention. Such is the case for establishing potency of vaccines, where Hendriksen, *et al.* (27) have shown that substantial reductions from conventional sample sizes are possible with no loss of ability to achieve a 95% confidence interval. A similar suggestion has been made by Shillaker, *et al.* (28) for skin sensitization tests and by Ennever and Rosenkranz (29) for carcinogenicity tests.

Artificial inflation of control group size: Frequently, investigators will ask for more animals than are required for the experiments they propose. We suspect that this happens because the investigator has not calculated the number of animals actually needed. When asked for a justification for the large number of animals, many of these investigators respond by decreasing the requested numbers. Usually the deliberations of the committee do not result in a change in requested numbers of animals. For example, in the most recent 109 protocols, 89% were approved for the number of animals originally requested. For 2/3 of the remainder, the investigator reduced the number of animals, and for 1/3 the investigator increased the number of animals as a result of questions asked by the committee. For those that reduced the number of animals, the average reduction was 33% (range 2-67%). Two investigators reduced the number of experimental treatments, but kept the number of animals the same.

Sometimes investigators propose using control groups that appear inordinately large. This can occur when the protocol involves sampling specific parameters at a series of times after a particular treatment. Control animals are then included for each time at which experimental animals will be used. This is often justified by the nature of the experimental design. However, when the control group does not compensate for variations in sample "time," the investigator may be trying to increase the apparent power of his experiment. As discussed in the previous section, this is not justified.

Protocols with several experiments in them will often request different sample sizes for each experiment. The committee must ask the investigator to explain why unequal sample sizes are to be used and, more to the point, how the smaller sample is justified if the larger one is actually needed or, conversely, why the larger sample is requested if the smaller one is sufficient. This problem occurs more often in experiments involving rodents, probably because they are most likely to involve large numbers of animals in several experiments.

Ways of Reducing the Number of Animals Used in Research

Nearly everyone agrees that it is desirable to reduce the number of animals used in research, but the effect on power of such reduction can be dramatic. Obviously, both committees and granting agencies must think carefully about the power of experiments before suggesting reductions in sample size (24). What are ways of reducing the number of animals used without reducing the power of the experiments?

Increase effect size: If the effect size, the minimum acceptable effect, can be increased then the size of the minimum sample needed to detect the effect will be reduced. One way to increase the effect size is to change the baseline against which the measurements are made. For example, if the effect to be measured is a change in blood glucose levels, then fasting the animals before the experiment may produce larger changes in these levels and increased effect size. Kempthorne (30) was able to magnify the effects of proteins on growth by reducing the nutrition level of the animals before the experiment. Towe and Mann (31) were able to detect an effect of strychnine in the cerebral cortex of cats simply by reducing the stimulus intensity. With supramaximal intensities the responses were saturated and an increase could not be detected. Other such maneuvers could be used in other situations.

This sort of manipulation may not always be appropriate. For example, dietary manipulations themselves could alter the response to an intervention, requiring an additional control group, or they may limit the generality of the conclusions drawn. The appropriateness of such manipulations depends upon the nature of the experiments. If the purpose is to see if an intervention *can* have *any* effect on a response, as in the case of the strychnine experiments of Towe and Mann (31), such manipulations are certainly warranted.

Reduce variability: Sample size may be reduced if the extraneous variability of the measurements in the sample can be reduced, a main purpose of designing experiments. One way to reduce variability is to increase the accuracy of the measurements. This may mean developing a new measurement tool. Another possibility is to use inbred animals (animals which are more alike genetically), littermate animals, or matched pairs of animals. Variability also can be reduced by using an animal as its own control.

Some experimental design elements themselves reduce or control for variability. The use of the randomized block design will control for variation. In addition, the primary purpose of analysis of covariance is to eliminate the effects of variables acting *with* the experimental variables. Even after all of these measures have been taken, there will still be variation in the data. This is the random variation in the measurements themselves and the variation among the individuals in the population of the parameter being

measured. These variations cannot be eliminated.

Wise use of control groups: The number of animals used in an experiment often can be reduced by careful use of controls. In some experiments, an animal may be used as its own control, thereby eliminating the need for a separate control group. This procedure has the added advantage of reducing variability by minimizing the effects of interanimal variation because paired tests may be used. Often the same sequence of observations is repeated under a variety of conditions in different experiments within the same protocol. Unless time is an important variable in such experiments, there is no necessity to duplicate control groups. This may give an added savings in animals.

Repeated samples from animals: With microanalysis techniques available, it is often possible to measure blood concentrations from very small samples. In these cases, it is possible to obtain samples repeatedly from the same animal over a period of time. This procedure will reduce the number of animals used by eliminating the need for separate groups to be sampled at each time. As in most procedures, this one has its limitations. The experimenter must either know that each sample has no effect on subsequent samples or he must be convinced that this is unimportant. For example, if blood is to be drawn from rodents by retroorbital bleeding, both animal welfare and scientific considerations dictate that this not be done too often. In addition, our committee requires that this procedure be done under anesthesia possibly introducing additional complications. The patency of indwelling catheters, especially when used in small animals, represents another limitation. Patency is frequently reduced with time, so the total duration of their usefulness may be limited.

When this kind of repeated sampling is done on the same animals, the statistical analysis must be done with care. The repeated measurements are not independent, because the value of each measurement is related to the measurements taken before. Shott (35) found that a major deficiency of papers published in two veterinary journals was treatment of dependent samples as if they were independent. With repeated measurements from the same animal, it is unreasonable to pool the samples as if they were equivalent. It is equally unreasonable to compare the sample at one time in a group of animals with the sample at another time in that same group using a 2-sample *t* test, a test which requires that the samples be independent. Under these circumstances, comparisons between experimental and control groups can be done legitimately.

Using replication or sequential testing: A sample size frequently proposed for protocols our committee has reviewed is 5-10 animals. Many investigators analyze their results in terms of the statistic *t*. For very large samples, the distribution of *t* approaches the normal distribution. As the sample size decreases, the peak of the distribution is depressed and the tails elevated, in short, it deviates more and more from normality. One of the assumptions of the test is that the data are, in fact, distributed normally. For very small samples, this is probably not true. There is no firm definition of what constitutes a "very small" sample, but a brief look at the *t* distribution will illustrate the difficulty with small samples. The 95% confidence interval is nearly 0.7 *t* units wider for a sample size of 5 than for a sample size of 1000. It is nearly 0.6 *t* units wider for a sample size of 5 than for a sample size of 35.

The power of the test decreases dramatically as the sample size is reduced from 35 to 5. Therefore, the probability of finding an effect if one exists becomes much smaller. Not only does the test become less reliable, but it also becomes less sensitive. The most common reason given for risking these errors with such small samples is that the appropriate number of animals would be too expensive or the experiments too time-consuming. Some investigators deal with this problem by replicating the experiment. That is, over the course of time, they make the same observations in several small samples instead of several large samples of animals. By pooling the results, they are able to get the pooled samples to an appropriate size. However, they incur the penalty of having to assume that the experiments were done in exactly the same way each time. As Tversky and Kahneman (32) have noted, "regardless of one's confidence in the original finding, its credibility is surely enhanced by the replication" if the effect is in the same direction and of approximately the same magnitude and if the variances are approximately equal.

The usual sampling techniques implicitly assume that a sample of prespecified size will be taken and observations will be made on the entire sample, regardless of whether all observations are actually needed to reach a decision about the hypothesis. Using a sequential test (33), it may be possible to reduce the sample size by as much as 50%. Sample units are selected randomly one at a time from the population. After each observation, the experimenter must decide whether or not to reject the null hypothesis or obtain another observation. (The decision is based upon the ratio of the probability function for the test under the null hypothesis , H_0, to that under the alternative hypothesis, H_1.) The α and β probabilities must, of course, be set in advance of the experiment. In this way, the experiment can be terminated before the usual preset number of animals has been used. Of course, the probability of detecting any event is greater if you're looking for it. This means that ordinary sampling statistics may not apply to serial tests. The interested reader should consult one of the many available sources for the methods of calculating probabilities and examples of application of these techniques (33,34,35).

Using one-tailed rather than two-tailed tests: Usually, an experimenter does not specify the direction of the difference between experimental and control groups. That is, he will accept a difference with the experimental measurements either greater than or less than the control measurements. This requires the use of a two-tailed test. However, if the experimenter is willing to specify the direction of the expected difference, then a one-tailed test may be used. The danger of using such a test is that it precludes finding results in the direction opposite to that expected. In some kinds of experiments, the opposite result may be of no interest, as in the case of studies of some potential therapeutic agent for depression. The investigator is not likely to be interested in an agent that increases depression.

A one-tailed test with an α probability of 0.05 will have the same power as a two-tailed test with an α probability of 0.10 provided the result is in the expected direction. If the result is in the other direction, the one-tailed test has no power at all. The result of the increase in power afforded by the one-tailed test is that a smaller sample can be used to achieve the same power as would be achieved by a two-tailed test with a larger sample. The investigator who elects this option must be aware that he pays the penalty of not being able to find a significant result in the direction opposite the expected one.

Using trend analysis to detect too small samples: If an experiment is performed with the preselected sample size, and the results are associated with an obtained p value greater than the α value selected *before* the experiment, most experimenters are inclined to simply not reject the null hypothesis. They will conclude that their intervention had no effect. There is still the possibility of a β error, of accepting the null hypothesis when it is false. It is always a good idea to carefully examine the data in any experiment. No experimenter should simply plug his data into a statistical program and confine his attention to the computer output. Examination of the data themselves may give an indication if a β error has occurred. For example, if the difference between the means of the experimental and control groups is not associated with a probability less than or equal to α, it may be that the sample size was too small. One way to detect this is to do a formal or informal trend analysis. If every experimental value was greater than every control value, it may be worthwhile to add more observations. On the other hand, if half were greater and half were smaller, adding more observations may well be a waste of animals.

Replacement of animals: The most effective way to reduce animal numbers would be not to use them at all. In some cases, it is possible to do studies on *in vitro* embryo cultures or cell cultures (36). For example, some questions regarding the infection of cells by viruses or effects of mitogens on ionic channels in lymphocytes can be answered effectively using cultured cells. This, by itself, *may not* reduce animal numbers unless cell-lines can be maintained and the number of available cells expanded because animals are the ultimate source of the cells. But not all research questions can be answered by use of cultures. Questions about the metastatic process, tissue or organ transplantation, brain mapping, behavior and many other processes cannot be answered at present using cultured cells. On the other hand, some questions regarding hematopoietic stem cells in peripheral blood or bone marrow, for example, can be answered using human cells and tissues which are available clinically, but not all such questions are answerable this way. If the cells must be introduced to a host and tissues collected from the host, humans cannot be used.

Computer or other types of models are commonly suggested to replace animals in research. In fact, some models may well prove to be useful for this purpose (37). But any model is only as good as the data and concepts on which it is based. We rarely understand any living system well enough to develop all-encompassing and effective models. It is clear that development of such a model would not eliminate the need for animals, but only reduce the number needed. Rosenkranz and Klopman (37) correctly emphasize that animal testing will always be needed from time-to-time to verify the continued validity of the model.

Conclusions

Committees charged with evaluation of experimental protocols must, by law, be constituted by both professional research workers and lay persons. Even the professional scientist may not be familiar with the statistical considerations outlined briefly here. It is too much to ask the lay representatives to make decisions based largely on such considerations. Nevertheless, the committee must make decisions about the protocols, including decisions about the appropriateness of the number of animals requested. It is clear to us that such decisions are not simple and seldom made upon statistical grounds alone. Members of committees must view the entire experiment in making decisions about animal numbers. We hope that we have highlighted some of the issues that have appeared in our consideration of a large number of protocols.

The subject of scientific merit has been purposely avoided here; it is a thorny issue that is best treated independently. However, we would be remiss in not pointing out that an experiment that is a waste of time is also a waste of animals.

Acknowledgments

The authors wish to thank Drs. Francis Clark and Roger Koehler and Ms. Sally Mann for their comments on earlier versions of this manuscript. They also wish to thank members of the committee for their diligent work in support of animal welfare.

References

1. Russell WMS, Burch RL. *The Principles of Humane Experimental Technique.* Methuen, London, 1959.

2. U.S. Congress. *The Animal Welfare Act.* 1966 PL 89-544, Vol. 7, U.S. Code 2131-2156. Amended December 17, 1985.

3. Office for Protection from Research Risks. *Public Health Service Policy on Humane Use of Laboratory Animals.* NIH, Bethesda, MD, 1986.

4. Part IV. Department of Agriculture. Animal and Plant Health Inspection Service. 9 CFR Parts 1,2, and 3 Animal Welfare; Final Rules. *Federal Register* 1989;54(168):36153.

5. Sokal RR, Rohlf FJ. *Biometry. The Principles and Practice of Statistics in Biological Research.* W.H. Freeman, San Francisco, 1969.

6. Altman DG, Gore SM, Gardner MJ, *et al.* Statistical guidelines for contributors to medical journals. *Brit Med J* 1983;286:1489-93.

7. Gardner MJ, Machin D, Campbell MJ. Use of check lists in assessing the statistical content on medical studies. *Brit Med J* 1986;292:810-12.

8. Berry G. Statistical guide-lines and statistical guidance. *Med J Austral* 1987;146:408-09.

9. Beyer WH. *Handbook of Tables for Probability and Statistics, 2nd Edition.* CRC Press, Boca Raton, 1968.

10. Aleong J, Bartlett DE. Improved graphs for calculating sample sizes when comparing two independent binomial distributions. *Biometrics* 1979;35:875-81.

11. Fleiss JL. *Statistical Methods for Rates and Proportions, 2nd Edition.* John Wiley & Sons, New York, 1981.

12. Glantz SA. *Primer of Biostatistics, 2nd Edition.* McGraw-Hill, New York, 1987.

13. Kastenbaum MA, Hoel DG, Bowman KO. Sample size requirements. *Biometrika* 1970; 57:421-30.

14. Schlesselman JJ. Statistics in veterinary research. *J Amer Vet Med Assoc* 1982;187:138-41.

15. Cohen J. *Statistical Power Analysis for the Behavioral Sciences, 2nd Edition.* Lawrence Erlbaum Associates, Hillsdale, NJ, 1988.

16. Erb HN. A statistical approach for calculating the minimum number of animals needed in research. *ILAR News* 1990;32(1):11-16.

17. Keppel G. Sensitivity of experimental designs. *Design and Analysis. A Researcher's Handbook.* Prentice-Hall, Inc., Englewood Cliffs, NJ, 1973;521-46.

18. Freiman JA, Chalmers TC, Smith H, *et al.* The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. *New Engl J Med* 1978;299:690-94.

19. Freiman JA, Chalmers TC, Smith H, *et al.* The importance of beta, the type II error, and sample size in the design and interpretation of the randomized controlled trial. Survey of 71 "negative" trials. *In* Bailar JC III, Mosteller F, eds. *Medical Uses of Statistics.* New Engl J Med Books, Waltham, MA, 1986; 289-304.

20. Overall JE. Classical statistical hypothesis testing within the context of Bayesian theory. *Psychol Bull* 1964;61:286-302.

21. Altman DG. Statistics and ethics in medical research: III How large a sample? *Brit Med J* 1980;281:1336-38.

22. Neyman J, Tokarska B. Errors of the second kind in testing "Student's" hypothesis. *J Amer Stat Assoc* 1936;31:318-26.

23. Kraemer HC. Sample size: When is enough enough? *Amer J Med Sci* 1988;296(5):360-63.

24. McCance I. The number of animals. *NIPS* 1989; 4:172-76.

25. Rowan AN. Research protocol design and laboratory animal research. *Invest Radiol* 1987;22:615-17.

26. Newton CM. Biostatistical and biomathematical methods in efficient animal experimentation. *The Future of Animals, Cells, Models, and Systems in Research, Development, Education and Testing.* National Academy of Sciences, Washington, DC, 1977;152-69.

27. Hendriksen CFM, Gun JW v d, Marsman FR, *et al.* The effects of reductions in the numbers of animals used for the potency assay of diphtheria and tetanus components of absorbed vaccine by the method of European Pharmacopoeia. *J Biol Stand* 1987;15:353-62.

28. Shillaker RO, Bell GM, Hodgson JT, *et al.* Guinea pig maximisation test for skin sensitisation: The use of fewer animals. *Arch Toxicol* 1989:63:283-88.

29. Ennever FK, Rosenkranz HS. Methodologies for interpretation of short-term test results which may allow reduction in the use of animals in carcinogenicity testing. *Toxicol Indust Health* 1988;4:137-49.

30. Kempthorne O. *Design and Analysis of Experiments.* Wiley, New York, 1952;179.

31. Towe AL, Mann MD. Effect of strychnine on the primary evoked response and on the corticofugal reflex discharge. *Exp Neurol* 1973;39:395-413.

32. Tversky A, Kahneman D. Belief in the law of small numbers. *Psychol Bull* 1971;76:105-10.

33. Ostle B. *Statistics in Research.* Iowa State University Press, Ames, IA, 1963.

34. Armitage P. *Sequential Medical Trials.* Blackwell Scientific Publications, Oxford, 1960.

35. Shott S. Statistics in veterinary research. *J Amer Vet Med Assoc* 1985;187:138-41.

36. Marazzi A, Ruffieux C, Randriamiharisoa A. Robust regression in biological assay: Application to the evaluation of alternative experimental techniques. *Experientia* 1988;44:857-73.

37. Rosenkranz HS, Klopman G. CASE, the computer-automated structure evaluation system, as an alternative to extensive animal testing. *Toxicol Indust Health* 1988;4:533-40.

38. Lachin JM. Sample size determinations for r x c comparative trials. *Biometrics* 1977;33:315-24.

39. Lakatos E. Sample sizes based on the log-rank statistic in complex clinical trial. *Biometrics* 1988;44:229-41.

40. Gatsonis C, Sampson AR. Multiple correlation: Exact power and sample size calculations. *Psychol Bull* 1989;106(3):516-24.

41. Gordon I. Sample size estimation in occupational mortality studies with use of confidence interval theory. *Amer J Epidemiol* 1987;125(1):158-62.

42. Satten GA, Kupper LL. Sample size requirements for interval estimation of the odds ratio. *Amer J Epidemiol* 1990;131:177-84.

43. Day SJ, Graham DF. Sample size and power for comparing two or more treatment groups in clinical trials. *Brit Med J* 1989;299:663-65.

44. Bach LA, Sharpe K. Sample size for clinical and biological research. *Aust NZ J Med* 1989;19:64-68.

45. Arkin CF, Wachtel MS. How many patients are necessary to assess test performance? *JAMA* 1990;263(2):275-78.

46. Mould RF. Clinical trial design in cancer. *Clin Radiol* 1979;30:371-81.

47. Beal SL. Sample size determination for confidence intervals on the population mean and on the difference between two population means. *Biometrics* 1989;45:969-77.

48. Feuer EJ, Kessler LG. Test statistic and sample size for a two-sample McNemar test. *Biometrics* 1989;45(2):629-36.

49. Fleiss JL, Levin B. Sample size determination in studies with matched pairs. *J Clin Epidemiol* 1988;41:727-30.

Footnotes

[1]The PHS and USDA use the term IACUC to designate the committee charged with local enforcement of animal welfare rules. In this paper, the term *committee* will be used to designate this body.

[2]The term power, in strict usage, applies to the statistical test, not to the experiment. McCance (24) has pointed out that its application to the experiment is justified because of the inextricable link between the experiment and the test of significance it employs.

[3]Our committee uses the term *protocol* to refer to a proposal to use animals in a research or teaching project. These protocols must follow a specified format. They are not proposals for funding; rather they concentrate on important animal welfare considerations (e.g., anesthesia, surgery, analgesia, euthanasia) as well as on research considerations (e.g., scientific justification, experimental design, and numbers of animals per test group).